KIOXIA

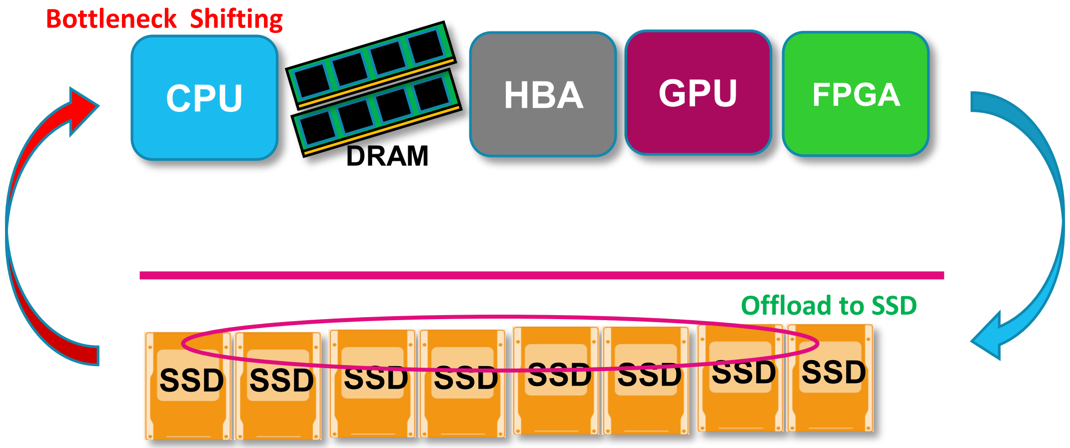
RAID Offload and Its Application

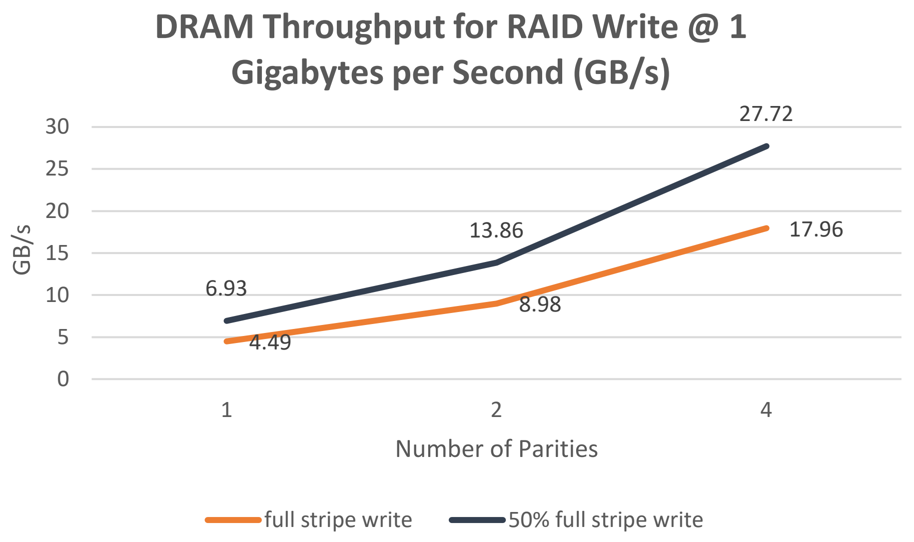
Aug 6, 2024

By Chandra Nelogal, Dell®
Devesh Rai, KIOXIA

RAID Offload Introduction

- NVMe™ SSDs performance improvement is continuously shifting the bottlenecks to applications
- Industry is addressing the problem in different ways
  - Hardware - RAID host bus adapters (HBA), data processing units (DPU), field programmable gate array (FPGA), configurable spatial accelerator (CSA)
  - Software - Costly CPU cores and dedicated DRAM
- **KIOXIA proposes to offload:**
  - Parity compute and memory resources to SSD
- **Benefits**
  - Reducing system total cost of acquisition (TCA) and total cost of ownership (TCO)
  - Continue to leveraging existing RAID applications and fault management





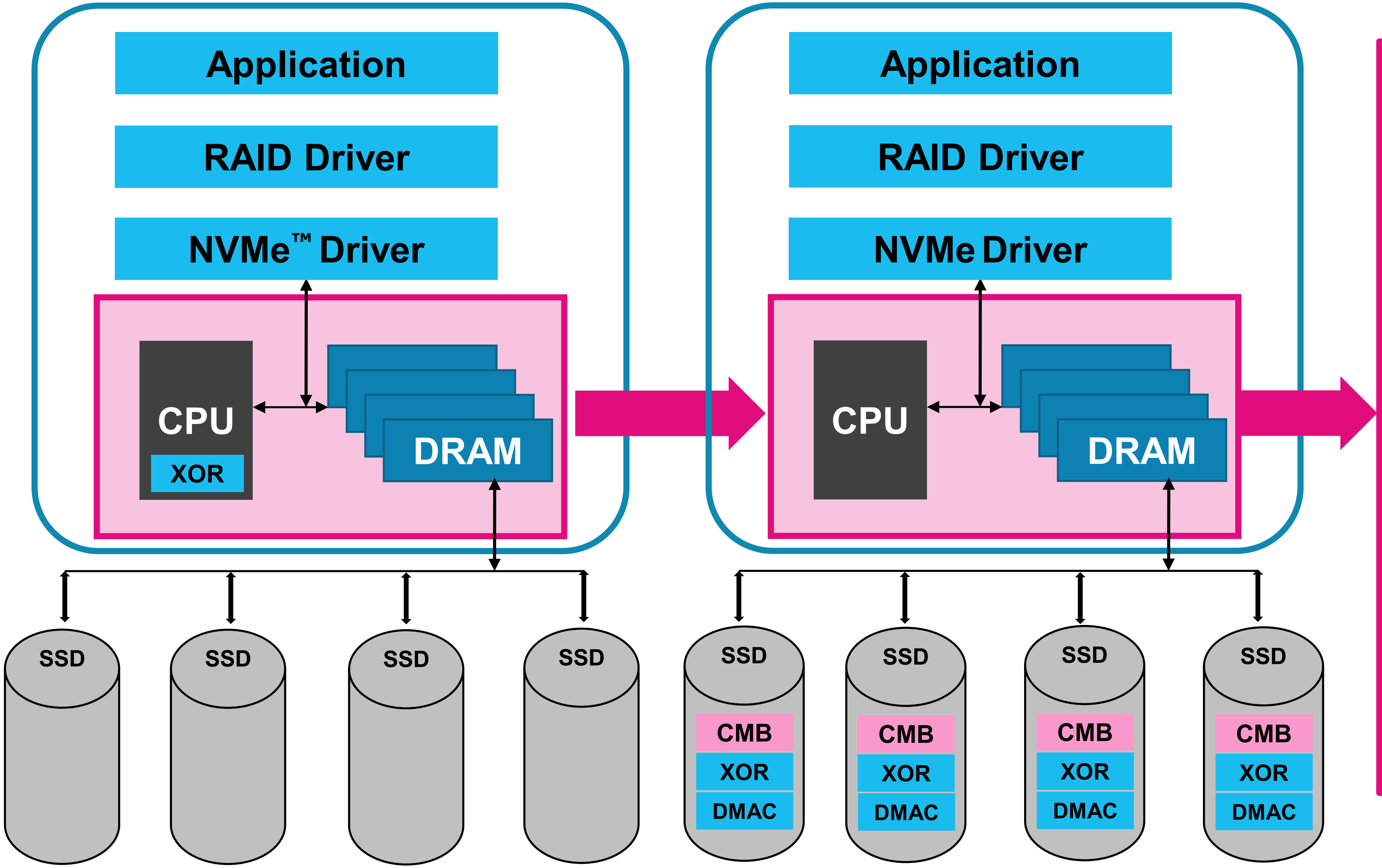
* Implementation specific

Images and product icons created by KIOXIA
Chart source: KIOXIA strategic marketing in-house testing and calculation

KIOXIA

The term "RAID" was invented by David Patterson, Garth A. Gibson, and Randy Katz at the University of California, Berkeley in 1987
Reference : https://www.usenix.org/system/files/conference/atc12/atc12-final181_0.pdf

KIOXIA RAID Offload Resource Utilization



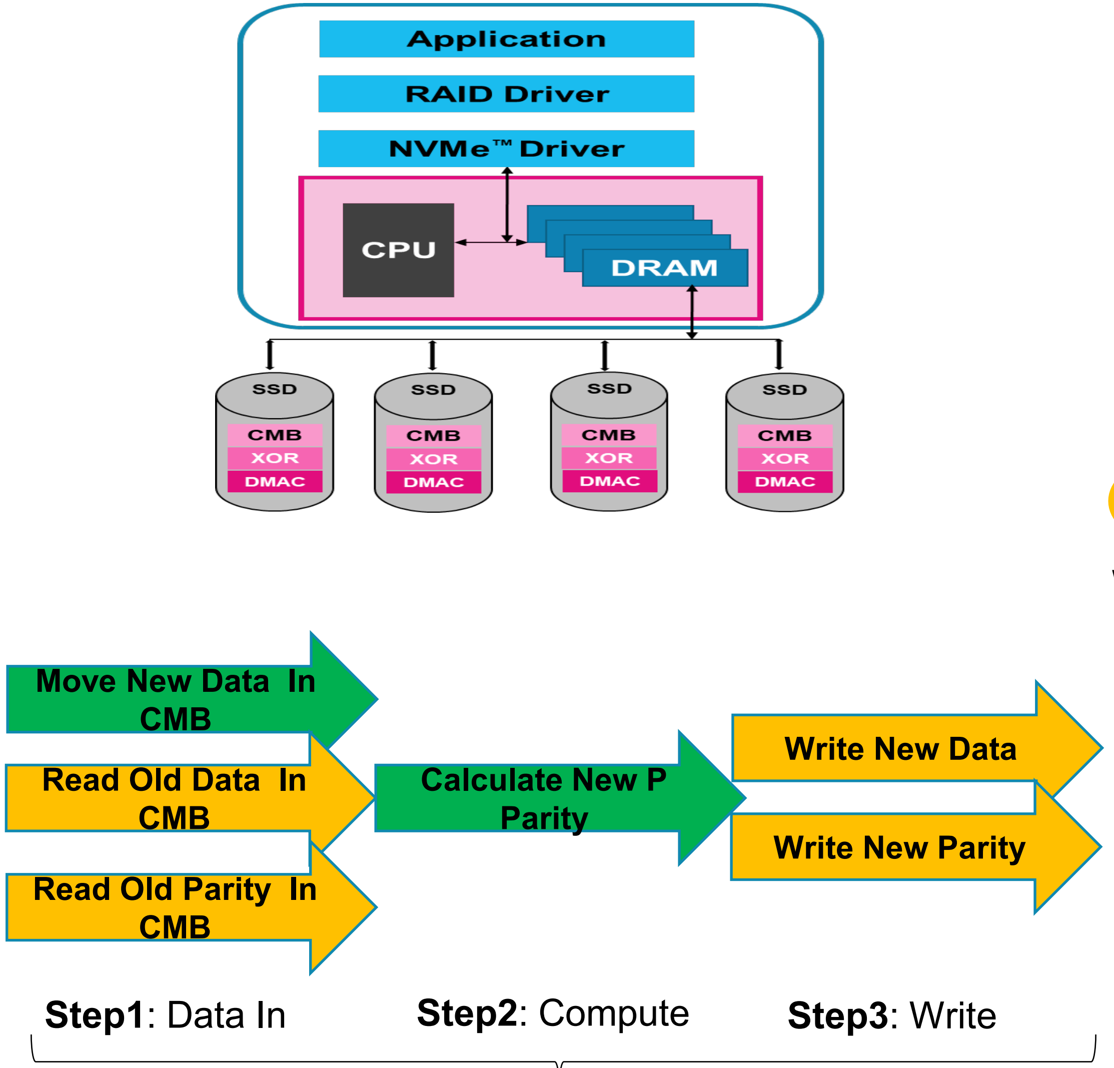
- Host managed standard based offload solution
- KIOXIA NVMe™ SSDs feature:
  - Controller memory buffers (CMB) – for DRAM offload
  - Exclusive OR (XOR) – up to 8 parity compute
  - Direct memory access controller (DMAC) – to place data in host address space (including remote CMB)
- Parallel compute and linear scaling

Graphics and product icons created by KIOXIA

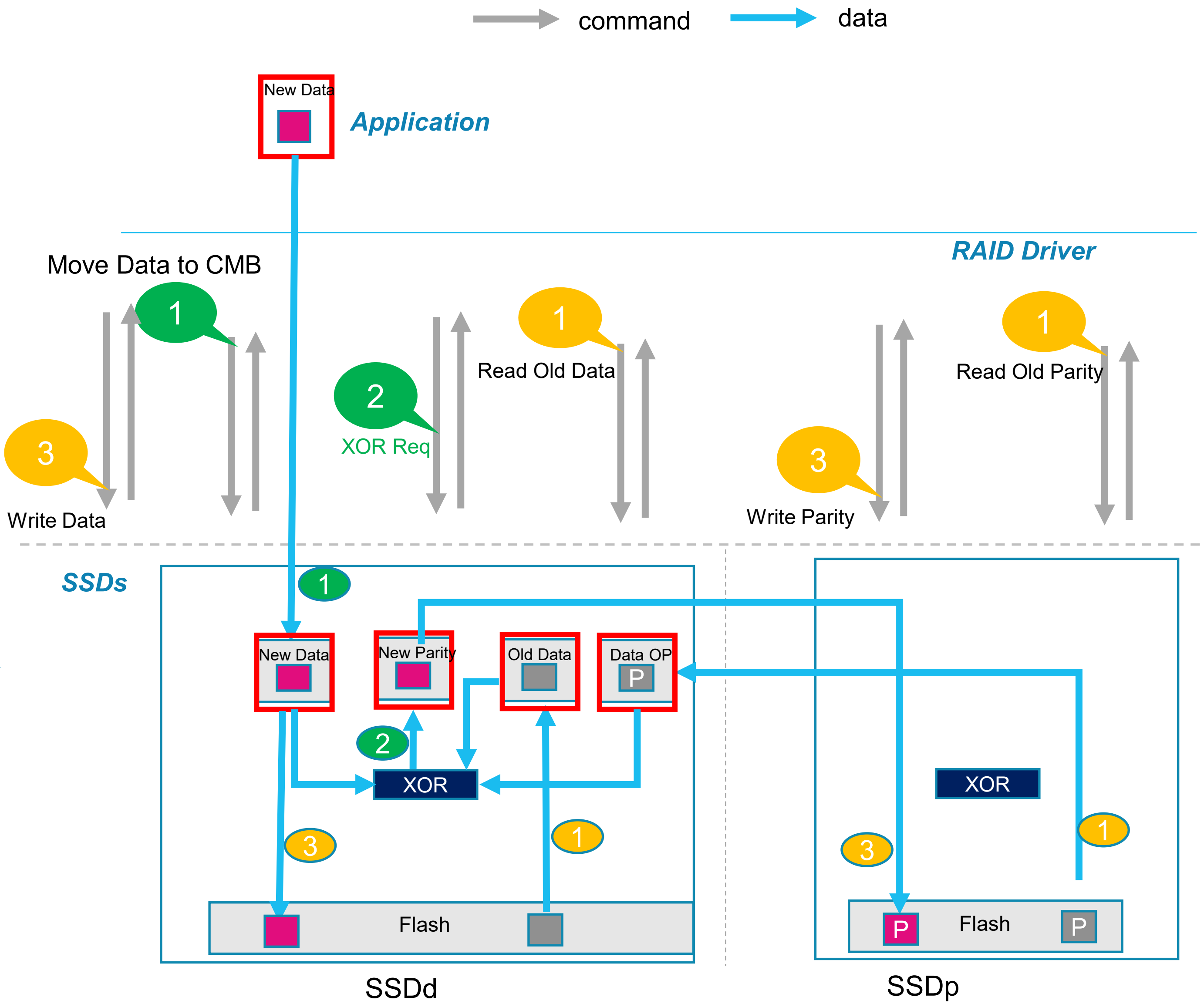
The term "RAID" was invented by David Patterson, Garth A. Gibson, and Randy Katz at the University of California, Berkeley in 1987

Reference : https://www.usenix.org/system/files/conference/atc12/atc12-final181_0.pdf

Command and Data Flow Example for RAID 5 Write

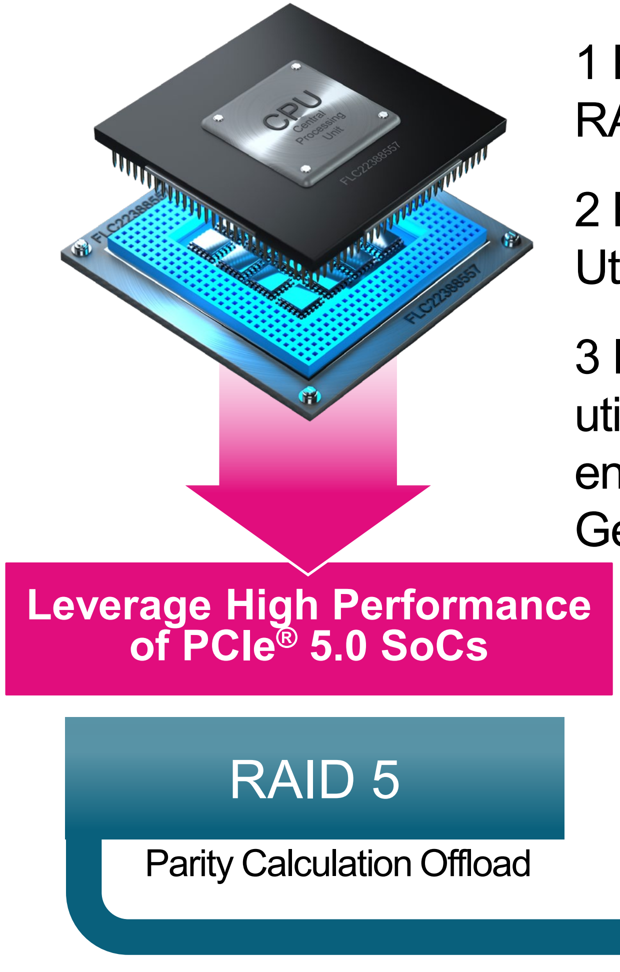


RAID 5 – Read/Modify/Write Sequence



Graphics, images and product icons created by KIOXIA
NVMe is a registered or unregistered mark of NVM Express, Inc. in the United States and other countries. PCI is a trademark of PCI-SIG

Proof of Concept (PoC) with mdraid5 and KIOXIA CM7 Series SSD



1 Reduce CPU workload for RAID computation

2 Reduce DRAM bandwidth Utilization

3 Improve host CPU utilization; contribute to energy efficiency for PCIe® Gen 5 server

Subject to Change Without Notice

RAID Offload : PoC Results (with KIOXIA CM7 & mdraid5)

System	KIOXIA CM7 Gen4 x4 – mdRAID 5#	RAID Offload	% Benefit
Number of SSDs	5	5	
Full Stripe Write 512 kibibytes (KiB)			
CPU Utilization	42	37	12% Reduction
DRAM Bandwidth in mebibytes (MiB/s)	3450	340	91% Reduction

IO workload: Flexible I/O tester (FIO) 512K Random Write @ 950 megabytes per second (MB/s)

System DELL® PowerEdge™ R650xs Xeon™ Gold 6338N 2.2GHz(2 Socket, 32 Cores) PCIe Gen4 , SSDs : 5xCM7 Gen4 (1.92TB)

The KIOXIA product images shown are a representation of the design model and not an accurate product depiction.

Image source:

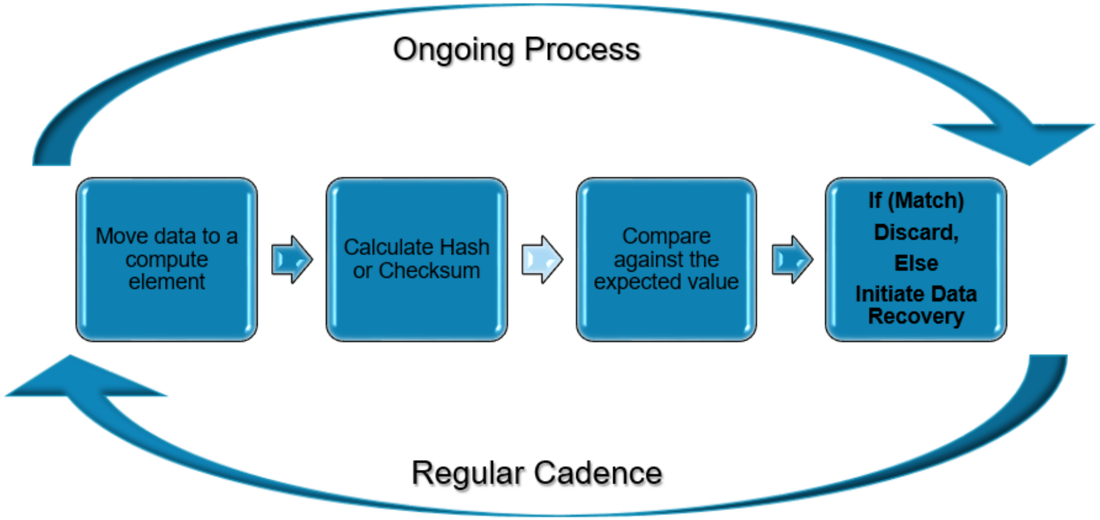
KIOXIA

NVMe is a registered or unregistered mark of NVM Express, Inc. in the United States and other countries. PCI is a trademark of PCI-SIG. Dell Technologies, Dell, PowerEdge are trademarks of Dell Inc. or its subsidiaries. Intel and Xeon are trademarks of Intel Corporation in the U.S. and/or other countries. . KIOXIA Corporation defines a megabyte (MB) as 1,000,000 bytes, a gigabyte (GB) as 1,000,000,000 bytes and a terabyte (TB) as 1,000,000,000,000 bytes. A computer operating system, however, reports storage capacity using powers of 2 for the definition of 1GB = 230 = 1,073,741,824 bytes

RAID Offload Use Cases

Data Scrubbing in Conventional Setup

- **Data Scrubbing: early detection and correction of errors**
- **Data Scrubbing technology: hash, checksum or RAID technology**

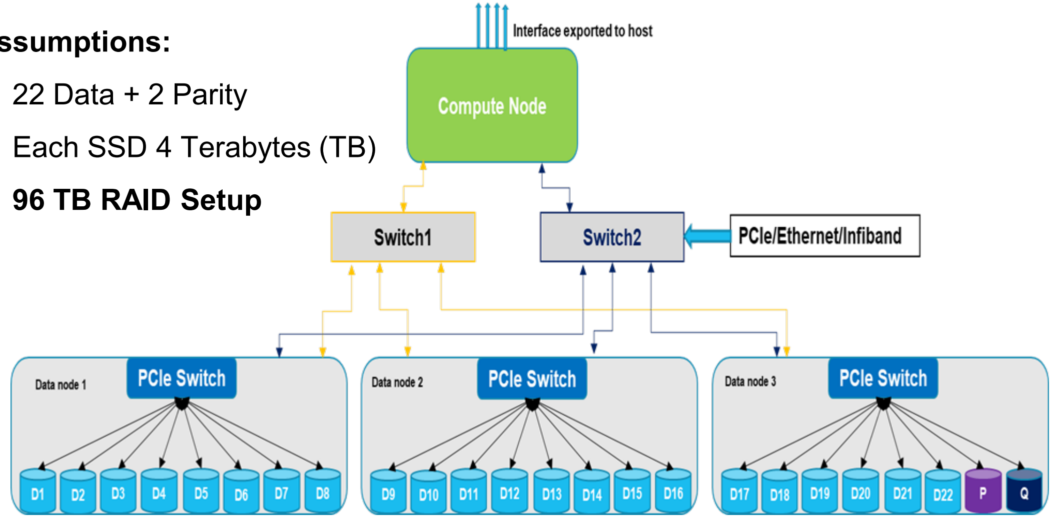


- **All data movement during scrubbing operation is an overhead penalty paid to ensure data integrity**

Graphics and product icons created by KIOXIA

Assumptions:

- 22 Data + 2 Parity
- Each SSD 4 Terabytes (TB)
- **96 TB RAID Setup**



Compute node performing disk scrubbing for one stripe using RAID

P + D1 + D2 + D3 + D4 …… + D22 = 0

Q + g1.D1 + g2.D2 + g3.D3 + ….. + g22.D22 = 0

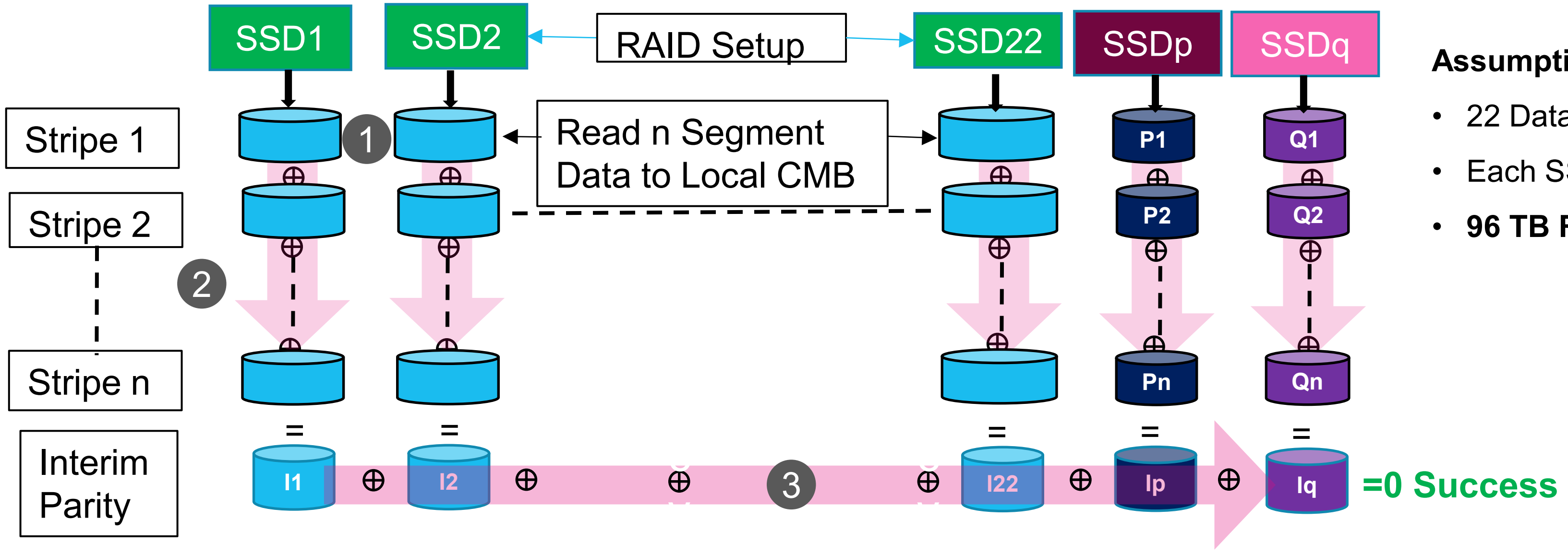
In above setup, 96TB data moves over PCIe®, network, CPU and 192TB through memory subsystem during each scrubbing cycle

Assumptions created by KIOXIA in-house engineering team

PCI is a trademark of PCI-SIG. KIOXIA Corporation defines a megabyte (MB) as 1,000,000 bytes, a gigabyte (GB) as 1,000,000,000 bytes and a terabyte (TB) as 1,000,000,000,000 bytes. A computer operating system, however, reports storage capacity using powers of 2 for the definition of 1GB = 230 = 1,073,741,824 bytes and therefore shows less storage capacity. Available storage capacity (including examples of various media files) will vary based on file size, formatting, settings, software and operating system, and/or pre-installed software applications, or media content. Actual formatted capacity may vary.

Data Scrubbing using RAID/EC Offload

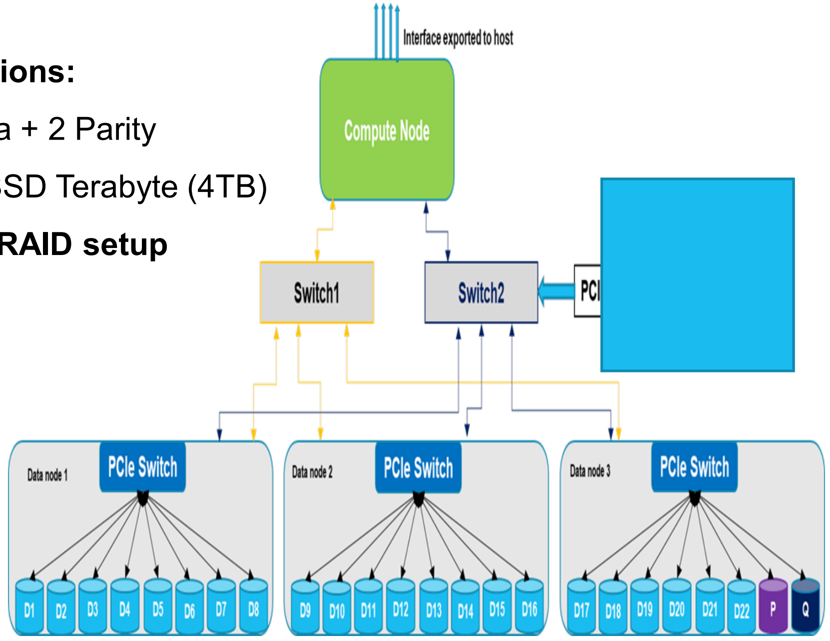
Subject to Change Without Notice



2. I1= D11 + g1.D11 + D21 + g2.D21 + ……+ Dn1 + gn.Dn1
3. I1 + I2 + I3 + ……. + Ip + Iq == 0 ➔ **Success**

Assumptions:

- 22 Data + 2 Parity
- Each SSD Terabyte (4TB)
- **96 TB RAID setup**

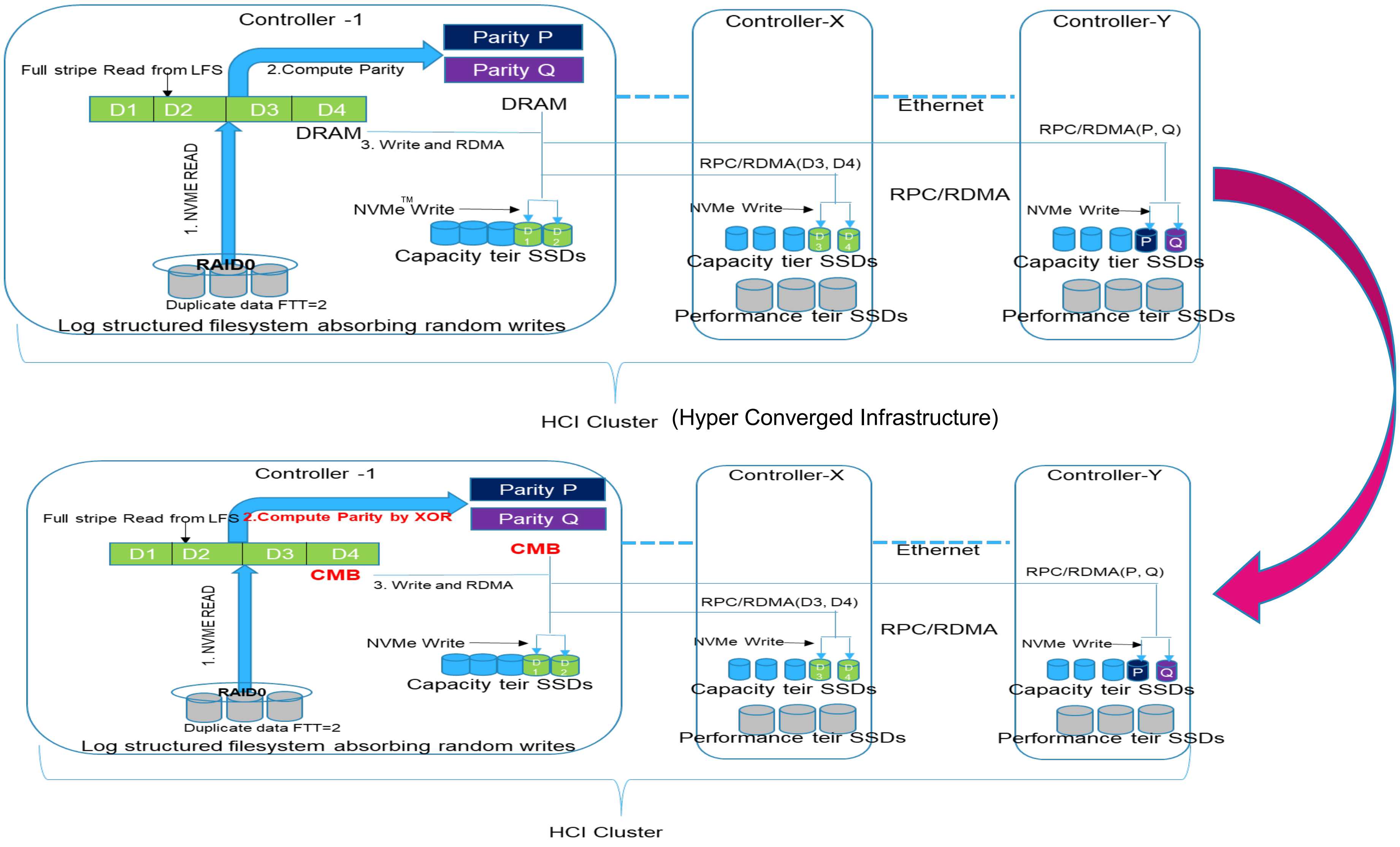


- ❖ Using 3 step process, ~99% data movement can be reduced
- ❖ No data passes through CPU and DRAM on compute node
- ❖ For n stripes, only one stripe moves over network and PCIe®
- ❖ Data scrubbing proof of concept data shown in table is for 9 SSD

Resource Utilization	Offload Disabled	Offload Enabled
Scrubbing time	129s	91s
DRAM Bandwidth	10.24 GB/s	1.43 GB/s
Total CPU Utilization	99.5%	~70%
L3 Cache Misses	14.7M	4M
Total PCIe Write (MB/s)	3694 MB/s	159 MB/s

Graphics, product icons, and tables created by KIOXIA. Assumptions created by KIOXIA in-house engineering team

RAID Offload in Hyperconverged, Software Defined Storage (SDS)



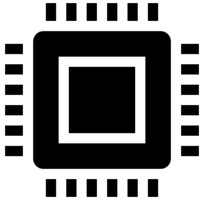
Controller saving system resources using RAID offload

Graphics and product icons created by KIOXIA
NVMe is a registered or unregistered mark of NVM Express, Inc. in the United States and other countries. PCI is a trademark of PCI-SIG

Future Possibilities: A Call to Action

- Offloading data scrubbing on to SSDs can significantly alleviate memory and network bandwidth bottlenecks and reduce data movement
  - Better resource utilization
- Easily adoptable in existing hyper converged infrastructure (HCI), RAID or similar solutions
- Dell® is collaborating with KIOXIA to standardize this technology

Standards Based



Host Controlled

Hardware Accelerators (Memory, Compute, DMAC)

Let's Collaborate! **Visit Booth# 307 for Demo**

KIOXIA

Example : Parity P and Q Generation Mechanism

RAID 6 P Parity: $P=D0 \oplus D1 \oplus D2 \ldots \oplus D31$

RAID 6 Q Parity: $Q=g0 \cdot D0 \oplus g1 \cdot D1 \oplus g2 \cdot D2 \ldots \oplus g31 \cdot D31$

1. g0 … g31 are Galois coefficient provided by host in XOR command.
2. D0 … D31 are per SSD data segment in a given full stripe
3. RAID application has option to calculate P or calculate Q or calculate both
4. Proposing up to 8 parity request in single command to support erasure code
5. Command structure may change during standardization process

Field for P parity command	Value
Source buffer address	D0, D1, ……D31
Galois coefficient for each buffer	1
Each source buffer length	16 Kilobytes (KB)
Output buffer address	P
Number of source buffer	32

Field for Q parity command	Value
Source buffer address	D0, D1, ……D31
Galois coefficient for each buffer	g0, g1, ……..g31
Each source buffer length	16 Kilobytes (KB)
Output buffer address	Q
Number of source buffer	32

Single XOR command calculating P and Q parity

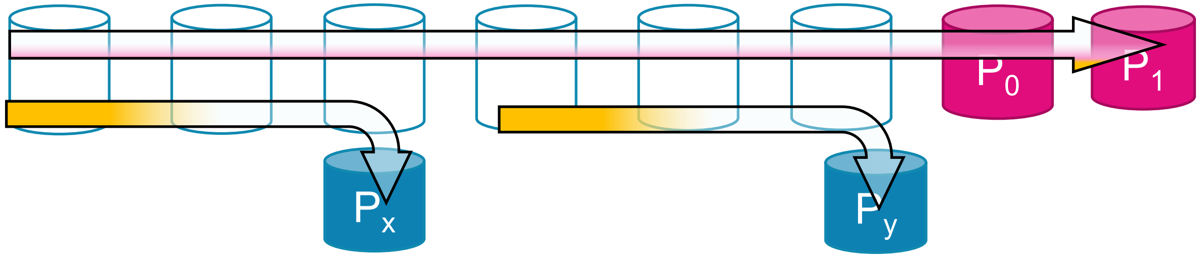
Tables: created by KIOXIA

KIOXIA Corporation defines a megabyte (MB) as 1,000,000 bytes, a gigabyte (GB) as 1,000,000,000 bytes and a terabyte (TB) as 1,000,000,000,000 bytes. A computer operating system, however, reports storage capacity using powers of 2 for the definition of 1GB = 230 = 1,073,741,824 bytes.

The term "RAID" was invented by David Patterson, Garth A. Gibson, and Randy Katz at the University of California, Berkeley in 1987

Reference : https://www.usenix.org/system/files/conference/atc12/atc12-final181_0.pdf

Use Case : Erasure Code (4 Parity Command)



- **Erasure Code** command for **4 parity** compute

Parity P_x	
Src buf	x0, x1, x2
Galois coefficient	1,1,1,1
Output buffer address	Px
Operation type	XOR

Parity P_y	
Src buf	y0,y1,y2
Galois generator	1,1,1,1
Output buffer address	Py
Operation type	XOR

Parity P_0	
Src buf	x0, x1, x2,y0,y1,y2
Galois coefffficient	$\alpha_0, \alpha_1, \alpha_2, \beta_0, \beta_1, \beta_2$
Output buffer address	P_0
Operation type	XOR

Parity P_1	
Src buf	x0, x1, x2,y0,y1,y2
Galois coefficient	$\alpha_0^2, \alpha_1^2, \alpha_2^2, \beta_0^2, \beta_1^2, \beta_2^2$
Output buffer address	P_1
Operation type	XOR

Graphics, product icons, and tables: Created by KIOXIA
The term "RAID" was invented by David Patterson, Garth A. Gibson, and Randy Katz at the University of California, Berkeley in 1987
Reference : https://www.usenix.org/system/files/conference/atc12/atc12-final181_0.pdf